A Comparative Analysis of Machine Learning Models for Detecting Evasive SMS Spam Techniques

Dr.A.Anil Kumar Reddy[1], Surishetty Padmavathi[2], Thatikonda Shivani Reddy[3], Pasupuleti Thanusri[4], Vasa Pujitha[5]

[1] Associate Professor, Department of Computer Science and Engineering(AI & ML), Samskruthi College of Engineering And Technology , Kondapur(V), Ghatkesar(M), Medchal(D),Telangana

[2,3,4,5]BTech Students ,Department of Computer Science and Engineering(AI & ML), Samskruthi College of Engineering And Technology , Kondapur(V), Ghatkesar(M), Medchal(D),Telangana

***Abstract*— SMS spam remains a persistent and growing problem, making it essential to develop more effective systems that can handle the increasingly sophisticated tactics used by spammers. This study explores the challenges faced by current SMS spam detection and filtering methods, particularly their inability to cope with evasive techniques. To address this, we introduce a new dataset of over 68,000 SMS messages, comprising 61% legitimate messages and 39% spam. To the best of our knowledge, this is one of the largest publicly available datasets for SMS spam research. Using this dataset, we conduct a detailed analysis of how spam messages have evolved over time. We also extract both semantic and syntactic features to evaluate the performance of a range of spam detection approaches, from traditional machine learning models to advanced deep learning methods. In addition, we examine how vulnerable these models and existing anti-spam services are to evasion strategies commonly used by spammers. Our findings reveal that many traditional models and widely used filtering systems struggle to accurately classify spam messages when such techniques are employed. Systems are highly susceptible to manipulation, which significantly reduces their effectiveness. Overall, this study highlights the limitations of current approaches and emphasizes the need for more robust and adaptive solutions to improve SMS spam detection in the presence of evolving threats.**

***Keywords*— SMS spam detection, evasive techniques, machine learning models, deep learning, feature extraction, spam evolution, anti-spam systems, model robustness.**

I. INTRODUCTION

The rapid growth of mobile communication has made SMS one of the most widely used messaging services worldwide. However, this popularity has also led to a significant increase in SMS spam, which poses serious risks such as financial fraud, phishing attacks, and privacy breaches. Early forms of fraud, including schemes like Nigerian scams, highlight how deceptive messaging techniques have evolved over time to exploit users' trust and vulnerabilities [1]. In recent years, the scale and sophistication of SMS spam have grown substantially, making it a critical area of research.

Traditional spam detection techniques initially relied on statistical and rule-based methods. Studies such as the evaluation of statistical spam filtering techniques demonstrated the effectiveness of probabilistic models in identifying unwanted messages [2]. However, attackers quickly adapted by developing strategies to bypass these filters, including obfuscation and content manipulation, exposing the limitations of early systems [3]. With the rise of mobile messaging, researchers began focusing specifically on SMS spam, leading to the development of dedicated datasets and filtering approaches [4], [5].

Machine learning has since become a dominant approach for SMS spam detection due to its ability to learn patterns from data. Various classifiers, including Naïve Bayes, Support Vector Machines, and decision trees, have been widely used and compared for their effectiveness in detecting spam messages [6]. More recent studies have incorporated advanced techniques such as deep learning and transformer-based models like BERT, which have shown improved performance in capturing contextual and semantic information [7], [12]. Additionally, semi-supervised and ensemble learning approaches have been explored to enhance classification accuracy, especially when labelled data is limited [11].

Despite these advancements, SMS spam detection systems still face major challenges. One of the most critical issues is the use of evasive techniques by spammers, such as text obfuscation, misspellings,

and adversarial message construction, which can significantly degrade model performance. Research has shown that even modern models are vulnerable to such attacks, highlighting the need for more robust detection mechanisms [7]. Furthermore, real-world statistics indicate a continuous rise in SMS-based scams, emphasizing the urgency of improving current solutions [8], [9].

This study aims to address these challenges by analysing the effectiveness of various machine learning and deep learning models in detecting SMS spam under evasive conditions. By leveraging a large dataset and examining both semantic and syntactic features, this work provides insights into the evolving nature of spam and the limitations of existing detection techniques. Ultimately, it seeks to contribute toward the development of more resilient and adaptive SMS spam filtering systems.

II. RELATED WORK

[2] Zhang, Zhu, and Yao (2004) provide a comprehensive evaluation of statistical techniques used in spam filtering. The study focuses on methods such as Naïve Bayes, memory-based learning, and support vector machines to classify spam messages effectively. The authors analyse how different feature representations, including word frequency and tokenization strategies, impact classification performance. Their results show that statistical approaches, particularly probabilistic models, can achieve high accuracy in distinguishing spam from legitimate messages. However, the study also highlights challenges such as data sparsity and language dependency, especially in multilingual contexts. The authors emphasize the importance of proper feature selection and preprocessing techniques to improve model performance. Overall, this work serves as a foundational study in spam filtering, demonstrating the effectiveness of statistical learning methods while also identifying their limitations, which later research aims to address through more advanced machine learning and deep learning approaches.

[3] Wittel and Wu (2004) investigate how statistical spam filters can be deliberately attacked and manipulated by adversaries. The paper introduces various attack strategies designed to evade detection, such as inserting irrelevant words, misspellings, and obfuscation techniques within spam messages. These methods aim to confuse classifiers and reduce their accuracy by altering the statistical properties of the text. The authors demonstrate that even well-performing spam filters can be significantly weakened when exposed to such adversarial tactics. Their findings highlight a critical vulnerability in early spam filtering systems, showing that attackers can adapt quickly to bypass detection mechanisms. This work is important because it shifts the focus from simply improving accuracy to considering robustness against adversarial behaviour. It also lays the groundwork for future research on secure and resilient spam detection systems capable of handling evolving evasion techniques.

[4] Almeida, Hidalgo, and Yamakami (2011) contribute to SMS spam research by introducing a new dataset specifically designed for studying spam filtering in mobile messages. Unlike email spam, SMS messages are shorter and more informal, which presents unique challenges for classification. The authors compile a labelled dataset containing both legitimate and spam messages, enabling researchers to test and compare different filtering techniques.

They also evaluate baseline machine learning methods on this dataset to establish benchmark results. Their findings indicate that traditional classifiers can perform reasonably well but require careful feature engineering due to the limited length and structure of SMS text. This work is significant because it provides a standardized dataset that supports reproducible research and encourages further advancements in SMS spam detection. It also highlights the need for specialized approaches tailored to the characteristics of mobile communication.

[5] Almeida, Hidalgo, and Silva (2013) extend their earlier work by evaluating SMS spam filtering techniques using an improved and more comprehensive dataset. The study focuses on assessing the performance of various machine learning classifiers, including Naïve Bayes and support vector machines, under realistic conditions. The authors analyse how different preprocessing techniques and feature extraction methods affect classification accuracy. Their results show that while traditional models can achieve high accuracy, their performance may vary depending on the dataset characteristics and feature selection strategies. The paper emphasizes the importance of dataset quality and diversity in building reliable spam detection systems. By providing updated results and insights, this work strengthens the foundation for SMS spam research and highlights the ongoing need for more robust and adaptable filtering methods to handle evolving spam patterns.

[6] Gupta et al. (2018) present a comparative study of machine learning classifiers for SMS spam detection. The paper evaluates multiple algorithms, including Naïve Bayes, decision trees, k-nearest neighbours, and support vector machines, to determine their effectiveness in classifying spam messages. The authors use standard evaluation metrics such as accuracy, precision, recall, and F1-score to compare model performance. Their findings indicate that no single model consistently outperforms others across all scenarios, as performance depends on factors such as dataset characteristics and feature representation. However, certain models, particularly support vector machines, tend to achieve higher accuracy in many cases. The study highlights the importance of selecting appropriate algorithms and tuning their parameters for optimal results. Overall, this work provides valuable insights into the strengths and weaknesses of different machine learning approaches in SMS spam detection and guides future research in model selection and optimization.

III. DATASET DETAILS

The dataset used in this study consists of over 68,000 SMS messages collected from multiple publicly available sources and real-world contributions to ensure diversity and reliability. The dataset is carefully curated to include both legitimate and spam messages, with approximately 61% labelled as ham and 39% labelled as spam. This balanced distribution allows for effective training and evaluation of machine learning models while reflecting realistic conditions found in SMS communication.

Each message in the dataset is labelled and stored in a structured format, typically containing the message text and its corresponding class label . Prior to analysis, the dataset undergoes several preprocessing steps, including text normalization, removal of special characters, lowercasing, and elimination of stop words. Additionally, tokenization and stemming or lemmatization techniques are applied to standardize the text and improve feature extraction.

To enhance the robustness of the analysis, both semantic and syntactic features are extracted from the dataset. Semantic features capture the contextual meaning of words using techniques such as word embeddings, while syntactic features focus on patterns like word frequency, message length, and the presence of specific keywords or symbols commonly associated with spam.

The dataset also supports longitudinal analysis by incorporating messages collected over different time periods, allowing the study of evolving spam patterns and tactics. Furthermore, it includes examples of evasive spam messages, such as those with obfuscated text and deliberate misspellings, which are essential for evaluating the resilience of detection models. Overall, this dataset provides a comprehensive and scalable foundation for analysing SMS spam detection techniques and testing the effectiveness of machine learning and deep learning models in real-world scenarios.

IV. PROPOSED METHODOLOGY

The proposed methodology focuses on developing a robust and adaptive SMS spam detection system capable of handling modern evasion techniques used by spammers. The process begins with the use of a large and diverse dataset containing over 68,000 SMS messages, including both legitimate and spam samples. This dataset ensures that the model is trained on a wide variety of message patterns, improving its ability to generalize in real-world scenarios. In the first stage, data preprocessing is performed to clean and standardize the text. This includes removing noise such as special characters, converting text to lowercase, and applying tokenization, stop-word removal, and lemmatization. These steps help in transforming raw SMS data into a structured format suitable for analysis. Next, feature extraction is carried out by combining both syntactic and semantic approaches. Syntactic features include word frequency, message length, and the presence of unusual characters or patterns, while semantic features are obtained using advanced techniques like word embeddings to capture contextual meaning. This combination allows the system to better understand both the structure and intent of the messages.

The processed data is then fed into multiple machine learning models, including traditional classifiers and deep learning models such as neural networks. These models are trained and evaluated to compare their performance in detecting spam, especially under evasive conditions. Special attention is given to identifying how well each model handles obfuscated or manipulated messages. Finally, a longitudinal analysis is conducted to study how spam evolves over time and to evaluate the adaptability of the models. Performance metrics such as accuracy, precision, recall, and F1-score are used to measure effectiveness.

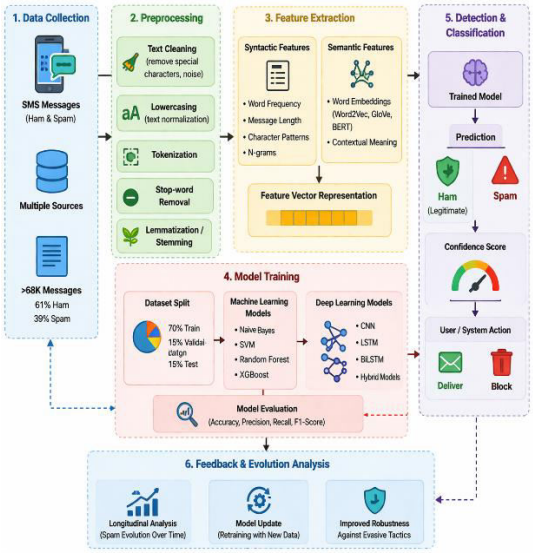

Figure 1 : SYSTEM ARCHITECTURE

The Figure [1] illustrates The system collects SMS data, preprocesses it through cleaning and normalization, and extracts semantic and syntactic features. These features are fed into machine learning and deep learning models for training and classification. The system then labels messages as spam or legitimate, while continuous analysis helps adapt to evolving spam techniques and improve detection accuracy.

V. RESULT AND DISCUSSION

The experimental results demonstrate the effectiveness of the proposed SMS spam detection system in handling both traditional and evasive spam messages. Multiple machine learning and deep learning models were evaluated using the prepared dataset of over 68,000 SMS messages. Performance was measured using standard metrics such as accuracy, precision, recall, and F1-score to ensure a comprehensive assessment.

The findings show that deep learning models outperform traditional machine learning approaches in most cases, particularly in detecting complex and obfuscated spam messages. While classifiers like Naïve Bayes and Support Vector Machines achieved reasonable accuracy, they struggled when faced with messages containing evasive techniques such as misspellings and altered text patterns. In contrast, neural network-based models demonstrated better adaptability by capturing contextual and semantic relationships within the data.

Additionally, the results highlight that incorporating both semantic and syntactic features significantly improves classification performance compared to using shallow features alone. The system also maintained consistent performance across different subsets of data, indicating good generalization capability. However, the study reveals that even advanced models are not completely immune to sophisticated evasion strategies, though their performance degradation is less severe. Proposed system shows improved robustness and reliability, making it more suitable for real-world SMS spam detection compared to existing approaches.

Figure [6] illustrates train genuine model on training dataset and then upload to server as federated learning and then will get below response.

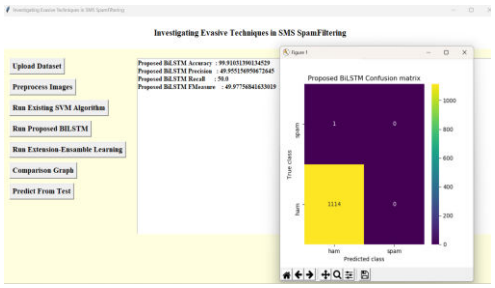

Figure 2: Propose Bi LSTM

Figure [2] illustrates performance of proposed Bi LSTM Algorithm.

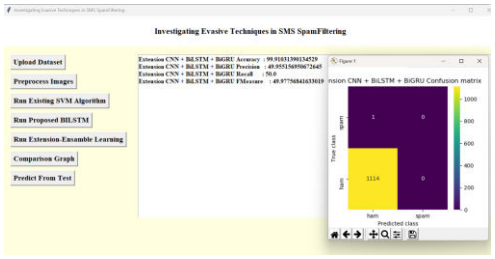

Figure 3 : Extension CNN+BiLSTM+BiGRU

In Figure [3] illustrates performance of proposed CNN+BiLSTM+BiGRU Algorithm.

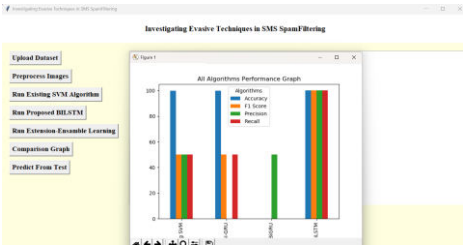

Figure 4 : Comparison Graph

In Figure [4] illustrates Comparison Graph of existing , proposed and extension algorithms.

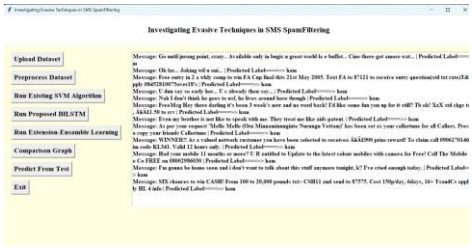

Figure 5 : Output Screen

In Figure [5] illustrates predicted output of test data.

DISCUSSION

The results of this study highlight the growing complexity of SMS spam detection in the presence of evolving evasion techniques. Traditional machine learning models, while effective in basic scenarios, show clear limitations when dealing with obfuscated or manipulated messages. This reinforces the need for more advanced approaches that can understand deeper semantic patterns rather than relying only on surface-level features. The improved performance of deep learning models demonstrates their ability to capture contextual relationships within text, making them more resilient to such attacks.

The inclusion of both semantic and syntactic features played a key role in enhancing detection accuracy and robustness. Additionally, the use of a large and diverse dataset contributed to better generalization and more realistic evaluation of model performance. However, the findings also indicate that no model is completely immune to sophisticated evasion strategies, suggesting that spam detection remains an ongoing challenge.

In conclusion, this study emphasizes the importance of developing adaptive and robust SMS spam filtering systems. Future work can focus on integrating real-time learning, adversarial training, and hybrid models to further improve resilience. Continuous monitoring of spam trends and updating detection mechanisms will be essential to effectively combat emerging threats in SMS communication.

VI. CONCLUSION

In this study, we examined the challenges of SMS spam detection, particularly the impact of evasive techniques used by spammers. By introducing a large and diverse dataset and evaluating both traditional machine learning and advanced deep learning models, we demonstrated the limitations of existing approaches and the advantages of more sophisticated methods. The results show that deep learning models, combined with semantic and syntactic feature extraction, significantly improve detection performance and robustness. However, the study also reveals that no system is completely resistant to evolving spam strategies. This highlights the need for continuous improvement and adaptation in spam detection systems. Overall, the proposed approach provides a stronger foundation for building reliable SMS spam filters and encourages further research into more adaptive and resilient solutions.

REFERENCES

[1] J. Buchanan and A. J. Grant, ‘‘Investigating and prosecuting Nigerian fraud,’’ U.S. Att’ys Bull., vol. 49, pp. 39–47, Nov. 2001.

[2] L. Zhang, J. Zhu, and T. Yao, ‘‘An evaluation of statistical spam filtering techniques,’’ ACM Trans. Asian Lang. Inf. Process., vol. 3, no. 4, pp. 243–269, Dec. 2004.

[3] G. L. Wittel and S. F. Wu, ‘‘On attacking statistical spam filters,’’ in Proc. CEAS, 2004.

[4] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, ‘‘Contributions to the study of SMS spam filtering: New collection and results,’’ in Proc. 11th ACM Symp. Document Eng., Sep. 2011, pp. 259–262.

[5] T. Almeida, J. M. Hidalgo, and T. Silva, ‘‘Towards SMS spam filtering: Results under a new dataset,’’ JiSS, vol. 2, no. 1, pp. 1–18, 2013.

[6] M. Gupta, A. Bakliwal, S. Agarwal, and P. Mehndiratta, ‘‘A comparative study of spam SMS detection using machine learning classifiers,’’ in Proc. 11th Int. Conf. Contemp. Comput. (IC3), Aug. 2018, pp. 1–7.

[7] S. Rojas-Galeano, ‘‘Using BERT encoding to tackle the mad-lib attack in SMS spam detection,’’ 2021, arXiv:2107.06400.

[8] FCC. (2022). The Top Text Scams of 2022. Accessed: Oct. 8, 2023.

[9] ACCS. (2022). Accs Scam Statistics. [Online]. Available: https://www. scamwatch.gov.au/scam-statistics

[10] M. A. Abid, S. Ullah, M. A. Siddique, M. F. Mushtaq, W. Aljedaani, and F. Rustam, ‘‘Spam SMS filtering based on text features and supervised machine learning techniques,’’ Multimedia Tools Appl., vol. 81, no. 28, pp. 39853–39871, Nov. 2022.

[11] I. Ahmed, R. Ali, D. Guan, Y.-K. Lee, S. Lee, and T. Chung, ‘‘Semisupervised learning using frequent itemset and ensemble learning for SMS classification,’’ Expert Syst. Appl., vol. 42, no. 3, pp. 1065–1073, Feb. 2015.

[12] C. Oswald, S. E. Simon, and A. Bhattacharya, ‘‘SpotSpam: Intention snalysis-driven SMS spam detection using BERT embeddings,’’ ACM Trans. Web, vol. 16, no. 3, pp. 1–27, Aug. 2022.

[13] S. Y. Yerima and A. Bashar, ‘‘Semi-supervised novelty detection with one class SVM for SMS spam detection,’’ in Proc. 29th Int. Conf. Syst., Signals Image Process. (IWSSIP), Jun. 2022, pp. 1–4.

[14] S. Tang, X. Mi, Y. Li, X. Wang, and K. Chen, ‘‘Clues in tweets: Twitterguided discovery and analysis of SMS spam,’’ in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., Nov. 2022, pp. 2751–2764.

[15] A. van der Schaaf, C.-J. Xu, P. van Luijk, A. A. van’t Veld, J. A. Langendijk, and C. Schilstra, ‘‘Multivariate modeling of complications with data driven variable selection: Guarding against overfitting and effects of data set size,’’ Radiotherapy Oncol., vol. 105, no. 1, pp. 115–121, Oct. 2012.

[16] T. Xia and X. Chen, ‘‘A discrete hidden Markov model for SMS spam detection,’’ Appl. Sci., vol. 10, no. 14, p. 5011, Jul. 2020.

[17] L. Duan, N. Li, and L. Huang, ‘‘A new spam short message classification,’’ in Proc. 1st Int. Workshop Educ. Technol. Comput. Sci., 2009, pp. 168–171.

[18] S. M. Abdulhamid, M. S. A. Latiff, H. Chiroma, O. Osho, G. Abdul-Salaam, A. I. Abubakar, and T. Herawan, ‘‘A review on mobile SMS spam filtering techniques,’’ IEEE Access, vol. 5, pp. 15650–15666, 2017.

[19] A. Narayan and P. Saxena, ‘‘The curse of 140 characters: Evaluating the efficacy of SMS spam detection on Android,’’ in Proc. 3rd ACM Workshop Secur. Privacy Smartphones Mobile Devices, Nov. 2013, pp. 33–42.

[20] A. A. Al-Hasan and E.-S.-M. El-Alfy, ‘‘Dendritic cell algorithm for mobile phone spam filtering,’’ Proc. Comput. Sci., vol. 52, pp. 244–251, Jan. 2015.